Text Mining

Programme course

6 credits

Text Mining

TDDE16

Valid from: 2017 Spring semester

Determined by
Board of Studies for Computer Science
and Media Technology

Date determined
2017-01-25

Main field of study
Information Technology, Computer Science and Engineering, Computer Science

Course level
Second cycle

Advancement level
A1X

Course offered for
- Computer Science and Software Engineering, M Sc in Engineering
- Computer Science and Engineering, M Sc in Engineering
- Information Technology, M Sc in Engineering
- Computer Science, Master's programme

Entry requirements
Note: Admission requirements for non-programme students usually also include admission requirements for the programme and threshold requirements for progression within the programme, or corresponding.

Prerequisites
Mathematical analysis; Linear Algebra; Probability and Statistics; Machine Learning; Basic programming.

Intended learning outcomes
The overall aim of the course is to provide an introduction to quantitative analysis of text, with special focus on applying machine learning methods to text documents. The student will learn all the main steps when working with text: i) efficient extraction of text, ii) natural language processing of the text in a form suitable for iii) statistical machine learning methods which are subsequently used for iv) text prediction.
After completing the course the student should be able to:

- use basic methods for information extraction and retrieval of textual data.
- apply text processing techniques to prepare documents for statistical modelling
- apply relevant machine learning models for analyzing textual data and correctly interpreting the results
- use machine learning models for text prediction
- evaluate the performance of machine learning models for textual data

Course content

Introduction and overview of quantitative text analysis and its applications. Information extraction. Web crawling. Information retrieval. Tf-idf. Vector space models. Text preprocessing. Bag of words. N-grams. Sparsity and smoothing for text. Document classification. Sentiment analysis. Model evaluation. Topic models.

Teaching and working methods

The course consists of lectures, computer laboratory work and an individual project. The lectures introduce concepts and theories that students then use in problem solving at the computer labs and in the project work.

Examination

PRA1	Project	3 credits	U, 3, 4, 5
LAB1	Laboratory exercises	3 credits	U, G

UPG1 consists of computer exercises that tests the students' ability to translate theoretical knowledge into practical problem solving in machine learning. UPG2 is an individual project where the student solves a real-world problem involving text. The project is documented and evaluated by a written project report.

Grades

Four-grade scale, LiU, U, 3, 4, 5

Department

Institutionen för datavetenskap

Director of Studies or equivalent

Ann-Charlotte Hallberg

Examiner

Marco Kuhlmann

Education components

Preliminary scheduled hours: 0 h
Recommended self-study hours: 160 h

Common rules

Regulations (apply to LiU in its entirety)

The university is a government agency whose operations are regulated by legislation and ordinances, which include the Higher Education Act and the Higher Education Ordinance. In addition to legislation and ordinances, operations are subject to several policy documents. The Linköping University rule book collects currently valid decisions of a regulatory nature taken by the university board, the vice-chancellor and faculty/department boards.

LiU's rule book for education at first-cycle and second-cycle levels is available at http://styrdokument.liu.se/Regelsamling/Innehall/Utbildning_pa_grund-_och_avancerad_niva.